Modeling Movements in Oil, Gold, Forex and Market Indices using Search Volume Index and Twitter Sentiments

Tushar Rao
Netaji Subhas Institute of Technology
Delhi, India
rao.tushar@nsitonline.in

Saket Srivastava
IIIT-D
Delhi, India
saket@iiitd.ac.in

ABSTRACT

Study of the forecasting models using large scale microblog discussions and the search behavior data can provide a good insight for better understanding the market movements. In this work we collected a dataset of 2 million tweets and search volume index (SVI from Google) for a period of June 2010 to September 2011. We model a set of comprehensive causative relationships over this dataset for various market securities like equity (Dow Jones Industrial Average-DJIA and NASDAQ-100), commodity markets (oil and gold) and Euro Forex rates. We also investigate the lagged and statistically causative relations of Twitter sentiments developed during active trading days and market inactive days in combination with the search behavior of public before any change in the prices/ indices. Our results show extent of lagged significance with high correlation value upto 0.82 between search volumes and gold price in USD. We find weekly accuracy in direction (up and down prediction) uptil 94.3% for DJIA and 90% for NASDAQ-100 with significant reduction in mean average percentage error for all the forecasting models.

Author Keywords

Stock market; sentiment analysis; Twitter; microblogging; social network analysis; oil; gold; forex

ACM Classification Keywords

H.1.m Information Systems: Miscellaneous

General Terms

Human Factors; Data Mining

INTRODUCTION

Most of the earlier works in computational finance comprise of *efficient market hypothesis* (EMH) that asserts market movements at the present level are a function of already existing news, whispers and the future valuation of dividends of a stock/ company [11, 16]. However research by Qian et al. shows markets are not fully efficient [18]. Behavioral finance challenges the very existence of efficient markets by placing the role of human sentiment and the social mood as vital part of investment decisions [17] and [22]. It challenges the EMH by adding the notion of human emotion and the macro-level mood play into investment decisions. For example at micro-level, consistently rising stocks is an indication of *selling to hold* the profits and perform subsequent portfolio adjustments. However surprising index trends are observed at macro-economic level. To further elaborate this point we will discuss the results about how bullish twitter sentiments can yield negative correlations with DJIA index (in other word bearish). This era of web technology is marked with high entropy of information spread as well as retrieval [3]. Earlier in late 90s before the spread of social web, information regarding commodities/currency rates and buy/sell sentiments took a long time (maybe even full day) to disseminate fully in the investor community. Also, the companies and markets took a long time (weeks or months) to calm market rumors, news or false information. This provides an opportunity to researchers to develop web mining platforms targeted towards mining relevant financial insights from the social media and web.

In this paper, we present a comprehensive study of relationships over wide range of market securities- commodities such as oil, gold, forex rates of Euro and equity markets such as DJIA and NASDAQ-100 with the dynamic features of the investor behavior as reflected in the opinions emerging on Twitter and trends in the search volumes. The summary of the study conducted is provided in the figure . Next, we discuss data collection and prior processing that explains the terminologies used in the market securities and social mood. Further we present the statistical techniques implemented and discuss the results and draw conclusions.

RELATED WORK

In the last decade several works have been done related to web mining of data (blogposts, discussion boards and news) [7], [9]. Some of them validate the significance of assessing behavioral changes in the public mood to track movements in stock markets. A recent work by Zhi et al. make use of search volume index (SVI) as indicator of investor demand for Russell 2000 stock for the time period from 2004 to 2008 [4]. Das and Chen made the initial attempts by using natural language processing algorithms classifying stock messages based on human trained samples [5]. However their result did not bring out statistically significant predictive relationships between message bullishness and the index returns [7, 9, 21, 1, 13, 12]. Eduardo et al. have used constraint graphs for each of the individual stock to explain effect of Twitter discussions [19].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WebSci'13, May 1 – May 5, 2013, Paris, France.
Copyright 2013 ACM 978-1-4503-1889-1 . . . $ 10.00.

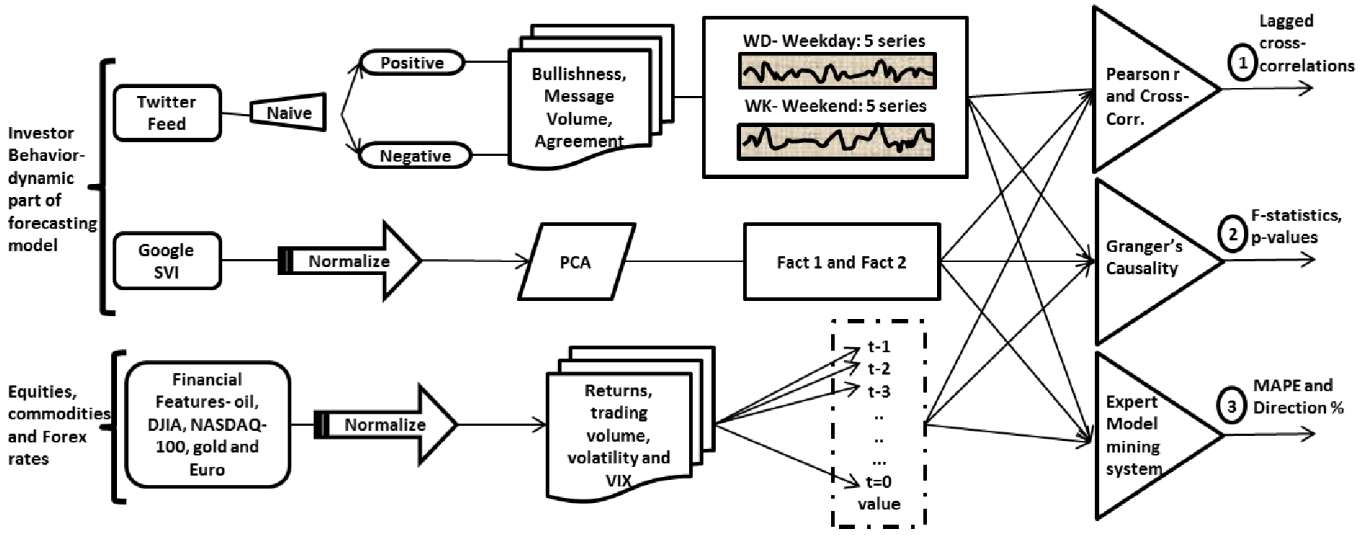


Figure 1. Flowchart of the proposed methodology showing the various phases of sentimental analysis beginning with SVI/ Tweet collection to stock future prediction. In the final phase three set of results have been presented:(1) Correlation results for twitter sentiments and stock prices for different companies (2) Granger's casuality analysis to prove that the stock prices are affected in the short term by Twitter sentiments (3) Using EMMS for quantitative comparison in stock market prediction using tweet features

Table 1. Comparison with prior work in sentiment analysis for predicting markets

Previous Approaches	Our Approach	Bollen et al. [1, 13] and Gilbert et al. [9]	Sprenger et al. [21]
Approach	Combining Twitter sentiment + Google search volumes	Mood of complete Twitter feed	Stock Discussion with ticker $ on Twitter
Dataset	2nd June 2010 to 13th Sept 2011- 1.9M tweets through search API	28th Feb 2008 to 19th Dec 2008, 9M tweets sampled as 1.5% of Twitter feed	1st Jan 2010 to 30th June 2010- 0.24M tweets
Techniques	Cross- Corr, GCA, Expert Model Mining System (EMMS)	SOFNN, Grangers and linear models	OLS Regression and Correlation

Gilbert and Karrie have used corpus from livejournal blogposts in assessing the bloggers sentiment with the dimensions of fear, anxiety and worry; making use of Monte Carlo simulation to reflect market movements in S&P 500 index [9]. Similar work is done by Bollen et al. who has used dimensions of Google- Profile of Mood States to reflect changes in closing price of DJIA [1]. Another work by Mao et al. covers effect of search volumes data in description with the preliminary sentiment indices of entire twitter feed on stock market movements of DJIA and volatility index of commodities like gold [14]. Zhang et al. also made have made use of dimensions in human behavior- fear and hope to show correlations with the stock market indicators [23]. However these approaches have been restricted to investor sentiment as derived from overall public mood and hence not flexible in terms explaining dynamics about individual stock index for companies. Sprengers et al. analyzed individual stocks for S&P 100 companies and tried correlating tweet features about discussions of the stock discussions about the particular companies containing the Ticker symbol [21]. From Table 1, we can see that earlier works in the area of behavioral finance were limited to profile of mood states and dimensions of public mood in context of investing. Primary objective of this work is to bring out a uniform model combining search volume behavior along with *how people are speaking and about what?* on Twitter and observe how severe or accurate these effects get over the increasing time lag. Use of bullishness, agreement and message volume although non-linearly dependent on each other, provides additional features to measure sentiment in a subjective manner and also provides better understanding of variable importance.

DATA COLLECTION AND PROCESSING

In this section, we discuss the collection of various financial data series used in this paper.

Tweets Extraction and Processing

Tweets are made accessible through a simple search of keywords(various market securities in our case) through an application programming interface (API)[1]. In this work, we have used tweets from period of 15 months and 10 days between June 2nd to 13th September 2011. During this period, by querying the Twitter search API for each of the market feature under study say Gold, Euro, Dow etc. we collected $1,964,044$ (by around 0.71M users) English language tweets. Each tweet record contains (a) tweet identifier,(b) date/time of submission(in GMT), (c) language and (d) text. Subsequently the stop words and punctuation are removed and the tweets are grouped for each day (which is the highest time precision window in this study, since we do not group tweets further based on hours/minutes).

[1]Twitter API is easily accessible at- https://dev.twitter.com/docs

Tweet Sentiment Extraction

In order to compute sentiment for any tweet we classify each incoming tweet everyday into *positive* or *negative* using naive classifier. For each day total number of positive tweets are aggregated as $Positive_{day}$ and total number of negative tweets as $Negative_{day}$. For sentiment analysis of tweets we have made use of lexicon/JSON API from Twittersentiment [2], a service provided by Stanford NLP research group [10]. Online classifier has made use of Naive Bayesian classification method, which is one of the successful and highly researched algorithms for natural language processing classification. It is known to give superior performance to other methods in context of tweets [10]. Because of the limitations of state of the art natural language processing algorithms, the accuracy of the classifier decreases tremendously when number of moods states (or number of classes) is taken higher. This decrease in sentiment accuracy affects the prediction accuracy as different in the rate of change of sentiment can be measured with significant precision. Naive Bayesian classification methods have high replicability and few arbitrary fine tuning elements.

In our dataset roughly 67.14% of the tweets are positive, while 32.86% of the tweets are negative for the market securities under study. This result indicates stock/ commodity discussions to be much more balanced in terms of agreement than chat and internet board messages where the ratio of positive to negative from earlier works ranges from 7:1 [6] to 5:1 [7]. Balanced distribution of stock discussion provides us with more confidence to study information content of the positive and negative dimensions of discussion about the stock prices on microblogs.

Feature Extraction and Aggregation

Further positive and negative tweets from each day are aggregated to make weekly time domain indicators which is the time period under study. We selected weekly domain over daily, bi-daily, bi- weekly or monthly as it is the most balanced time resolution to study the effect of investor behavior over model performance accuracy; keeping in-market monetization potential practically impeccable.

For every week, the value of the security (closing, volatility, volume, weekly returns for each index) is recorded every Friday at closing time of the market trading hours 21:00 UTC. To explore the relationships between weekly trading and also on the days when market remains closed (weekends, national holidays), we broadly focus on two domains of tweet sentiments- weekday indices and weekend indices (further referred as *WD* and *WK* respectively). We have carried forward the work of Antweiler et al. [7] for defining bullishness (B_t) for each time domain (time window is *WD* or *WK*) as given by equation 1:

$$B_t = \ln \frac{1 + {M_t}^{Positive}}{1 + {M_t}^{Negative}} \quad (1)$$

Where ${M_t}^{Positive}$ and ${M_t}^{Negative}$ represents the number of positive or negative tweets during a particular time period *WD* or *WK*. Logarithm of bullishness measures the share of surplus positive signals and also gives more weight to larger number of messages in the specific sentiment group (positive or negative). Message volume is simply defined as natural logarithm of total number of tweets per time domain for a specific security/index. And the agreement among positive and negative tweet messages is defined as:

$$A_t = 1 - \sqrt{(1 - \mid \frac{(M_t^{Positive} - M_t^{Negative})}{(M_t^{Positive} + M_t^{Negative})} \mid} \quad (2)$$

If all the tweet messages about a particular company are positive (bullish about a company stock) or negative (bearish about a company stock), agreement would be 1 in that case. Influence of silent tweets days in our study (trading days when no tweeting happens about particular company) is less than 0.1% which is significantly less than previous works [7, 21]. Every market index/ security thus have a total of 10 potentially causative time series from Twitter: positive *WD*, negative *WD*, bullishness *WD*, message volume *WD*, agreement *WD* and from previous weekend we have positive *WK*, negative *WK*, bullishness *WK*, message volume *WK* and agreement *WK*.

Search Volume Index

To generate search engine lexicon for each of the five securities under study- Oil, DJIA, NASDAQ-100, Gold and Euro; we start by collecting weekly search volumes for specific search terms related to respective sectors like- oil, GLD, Dow-30, nasdaq, oil price etc. as given in Table 2 from **Google Insights of Search** [3]. Google provides this open service to access the search volume data at weekly minimum frequency since January 2004. Next we also take into account the top recommended relevant search terms by Google insights of search, thus expanding the already existing group of search terms.

To further normalize and better understand the computational results, we apply dimension reduction technique of principle component analysis. We are able to reduce the number of variables (uptil 50 for oil) from search domain by combining similarly behaving time series to create completely uncorrelated co-independent factors- Fact 1 and Fact 2. Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components which reveals underlying structure that is responsible for maximum variance. As given in Appendix, Table 6, 9, 7, 8 and 5 gives the extracted factors by varimax rotation technique to produce orthogonal factors. To identify the factors that cause maximum variance in retweets, we have used Kaiser criterion in which the factors with eigen values greater than 1 are extracted.

Financial Market Data

Our analysis broadly covers five different sectors oil, DJIA , NASDAQ-100, gold and Euro. Most of the data, including all the VIX indices and Euro to USD fedex rates used for analysis are collected from econometrics data from Federal

[2]https://sites.google.com/site/twittersentimenthelp/

[3]http://www.google.com/insights/search/

Table 2. Google search Terms for 5 Securities

US Oil Funds	oil commodity, crude oil, oil etfs, curde oil price, oil futures, oil quotes, oil price per barrel, oil prices bloomberg, wti crude oil, oil prices, how much of oil is left, crude oil ticker + 50 more similar terms etc.
DJIA	djia, dow jones industrial average, dow jones, dow, s&p 500, Stock Market, stock message board
Nasdaq-100	nasdaq up, djia today, dow futures quote, futures quote, djia quote, nylc, bank of america dividends
Gold	buy gold, invest in gold US data, invest in gold worldwide, dollar to pound exchange rate, dollar to pound exchange
Euro	exchange rates converter,dollar euro exchange rate history,rupee exchange rate, oanda currency,rupee exchange,dollar rupee exchange rate,bloomberg live tv,eurusd

Reserve Bank of St. Louis [4]. Gold prices time series are downloaded from World Gold Council [5]. Weekly time series for US oil funds and weekly index movements in DJIA and NASDAQ-100 are extracted from Yahoo Finance! API[6].

The financial features (parameters) available from Yahoo finance under study are opening (O_t) and closing (C_t) value of the stock/index, highest (H_t), lowest (L_t) value and volume traded for the stock/index. In addition returns are defined as difference between the logarithm of closing values of the stock index between the week's Friday and previous week's Friday.

$$R_t = \{\ln Close_{(t)} - \ln Close_{(t-1)}\} \times 100 \qquad (3)$$

Trading volume is the logarithm of number of traded shares every week. We estimate weekly volatility based on intraday highs and lows using Garman and Klass volatility measures [8] given by the formula:

$$\sigma = \sqrt{\frac{1}{n}\sum \frac{1}{2}[\ln \frac{H_t}{L_t}]^2 - [2\ln 2 - 1][\ln \frac{C_t}{O_t}]^2} \qquad (4)$$

Further in this section we will discuss the various security indices in each of the sector under study.

Oil

In this study we have taken USO- United States Oil Fund, an exchange traded fund (ETF) that is one of the highly traded security and strongly tracks movements of light, sweet crude oil purchased and sold at NYSE Arca. We have extracted weekly closing values, volatility and volume parameters from the lexicon. In addition to this we have also taken CBOE OIL volatility index [7] (further referred as VIX) which is index measure of market's expectation of 30-day volatility of crude oil prices. Volatility is a measure for variation of price over a period of time, whereas VIX is measure of the market's expectation of stock market volatility over the next 30 day period. VIX is often used a parameter for deciding investments in Futures and Options markets.

[4]Federal Reserve Economic Data: http://research.stlouisfed.org/fred2/

[5]http://www.gold.org/investment/statistics/goldpricechart/

[6]http://finance.yahoo.com/

[7]http://www.cboe.com/micro/oilvix/introduction.aspx

DJIA

Its an aggregate of 30 highly traded and influential stock evenly distributed over all sectors. We have taken weekly returns, volatility and volume as parameters under the study. Further we have also extracted CBOE DJIA VIX which is indicative measure of fluctuation in 30-day future index sensitivities.

NASDAQ-100

Its an aggregate of the top 100 stocks from $NASDAQ$ exchange which indexes majority of the technological stocks in the market. For this as well we have taken weekly returns, volatility and volume as the parameters under study. In addition we also extracted CBOE NASDAQ-100 VIX which is indicative measure of 30-day ahead index movements.

Gold

We have taken price in US dollar (USD) as its the most traded currency for gold in the world to accurately represent search volumes in each country and related twitter buzz for the precious metal. Further we have extracted Gold ETF VIX as well from CBOE, as indicative of a month ahead fear-gauge in the price of the precious metal.

Euro

We have taken only two parameters- one Euro to USD (US dollar) conversion rates at closing of the market on Friday's eve for every week and other CBOE Euro ETF VIX as measure of 30-day market fear for the same.

STATISTICAL ANALYSIS AND RESULTS

In this section we begin statistical analysis and forecasting performance on each of the financial securities as discussed in the previous section from two dynamic components investor behavior comprising of 10 components from Twitter as discussed earlier and 1 or 2 principle factors from Google SVI as discussed in Appendix.

First we identify correlation patterns across various time series at different lagged intervals, further testing the causative relationships of SVI and tweet features on the market securities using econometric technique of Granger's Casuality Analysis. Then we make use of expert model mining system (EMMS) to propose and test the forecasting model and draw performance based conclusions.

Correlation and Cross-Correlation Analysis

We begin our study by identifying pairwise correlation metrics between 10 Twitter features for each security index given in previous section and the factors derived from SVI search factors as given in Appendix.

Technique

Once we obtain the pearson correlation coefficients, as an evaluation of the lagged response of relationships existing between financial features, Twitter sentiments and the search volumes; we compute cross-correlation at a lag of $\pm$ 7 week lag to show confidence and effectiveness in results. It also motivates us to look forward in making an accurate forecasting model by picking accurate regressor co-efficient.

For any two series $x = \{x_1, ..., x_n\}$ and $y = \{y_1, ..., y_n\}$, the cross correlation lag γ at lag k is defined as:

$$\gamma = \frac{\sum_i (x_{i+k} - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_{i+k} - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}} \quad (5)$$

In equation 5, $\overline{x}$ $\overline{y}$ are the mean sample values of x and y respectively. Cross-correlation function defined as short for ccf(x,y), is estimate of linear correlation between x_{t+k} and y_t, which means keeping the time series y stationary, we move the time series y backward to forward in time by a lag of k i.e. k= [-7,7] for lags of 7 weeks in positive and negative direction. Cross-correlation gives the measure of anticipated values of statistically significant relations in a time series x which can be made part of the forecasting models discussed ahead.

Results

The Figure 2 as heatmap along with Figure 3 as radar maps, represents summarized set of pearson correlation results for financial, Twitter and SVI time series after transformation to log scale. Corresponding red and blue sections in both the figures correspond to statistically significant relationships between the dependent variables.

For Twitter features we examine 5 series- positive, negative, bullishness (Bull), message volume (Msg Vol.) and agreement (Agrmnt) as two cases one as weekday(active market trading) and other as weekend (during market off days). We realize that the overall nature of relationship exhibit varying degree of association. But the clear trend that we observe is that market-off days don't carry high weights when compared to overall data available on comparison to market active days but its still significantly correlated and can be potentially exploited while designing the hedging strategies as discussed later in this paper. Weekday bullishness is one of the important feature out of all others to look out for any investment and show uniformly significant behavior in all the sectors with value of pearson 'r' as high as -0.73 for DJIA's weekly volatility. Another interesting trend that is observed is returns in both DJIA and NASDAQ-100 show negative relationship of varying strength with both positive and negative feeds indicating heavy discussion which is more sensitive to message volume on Twitter before fall in the index. But significantly valuable relationship of 0.593 correlation exists with the returns with our introduced feature term bullishness which is relative measure of positive to negative sentiment of investor community as explained in earlier section . NASDAQ doesn't carry any relationship with weekends Twitter discussions on account fast and dispersive behavior of news memes among tech-savvy investors of technological stocks whom are expected to be faster response to news. For volatility indices (VIX) for various securities shows significant negative relation with weekday agreement index which is vector distance between positive and negative discussion about any security as measure of accurate picture of about to happen turbulence/ perceived market risk in the coming weeks; except for DJIA which consists major 30 stocks only which are subjected to highly balanced consistent movements due to heavy trading activity across any time domain.

We find stronger correlation of the principle factors from SVI series uptil 0.826 for commodity funds like for oil, gold and Euro forex rates as compared to index movements of DJIA and NASDAQ-100; giving an impression that people tend to search more for commodity funds then stock equities indicating a better understanding of control heuristics of actual market movements from investor behavior. From Figure 2 we can see that the VIX is one of the highly correlated financial feature in all the 5 cases, thus maybe referred as a strong measure of investor behavior though *computational gauge of investor fear*. Another important significant relation that we observed for NASDAQ-100 and DJIA is the negative correlation with returns in contrast to positive correlation of volatility, volume and VIX; which is indicative of high search behavior being caused by fall in the index values, increasing more volume in trading making the index movements more volatile.

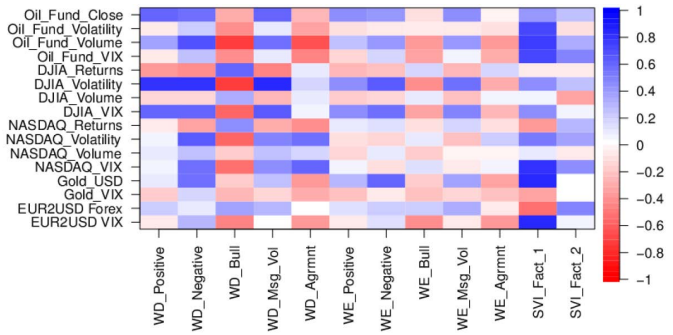
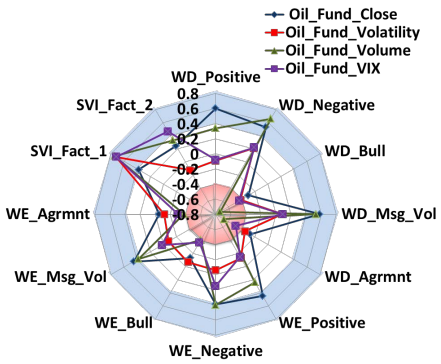


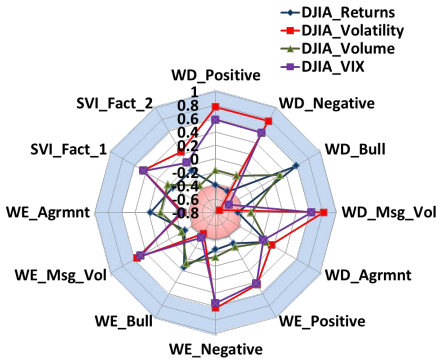
Figure 2. Heatmap showing pearson correlation coefficients between security indices vs features from Twitter and SVI factors. (Blue and red correspond to significant correlation values in Figure 3)

As we can see in the figure 4 (a), highest correlation is exhibited by oil VIX and SVI which is roughly balanced on both the sides indicating a bi-causative relation in both the directions. Similar observations can be seen for oil fund and oil VIX to Twitter message volume. Tweet message volume have stagnant low slope on the negative lag side which indicates surge in oil related discussion on Twitter consistently prior to actual hike in the price. For DJIA and NASDAQ-100 as observed from figure 4 (c) and (d), much balanced correlation factors can be observed for majority of the pairs in both the cases. However, for DJIA significant bend on the negative lag side is observed by volatility in the index for k=-1, indicating a fall of -0.8 correlation in tweet based bullishness atleast a week before the actual market trading. Similar effect is observed for search volumes uptil 4-5 weeks before the actual trading volume increases. NASDAQ-100's correlation activity doesn't give much insights into relationships between the features which maybe due to non- linear associations or significant relations hidden at smaller time domains frequencies as nature of tech-savvy investors of technological stocks. But we can see that bend on positive *k* lag side for volatility with search volumes for a week before and constant increase in bullishness prior *2* weeks before actual surge in volatility. However we leave this area for future exploration.

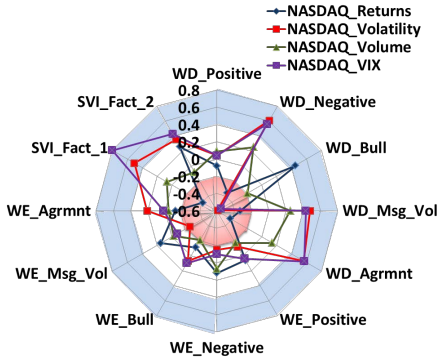
From figure 4 (b) and (e) we can see balanced correlation



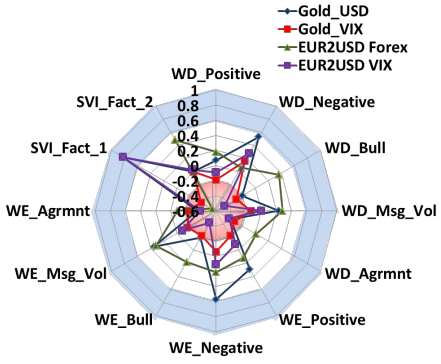
(a) Oil



(b) DJIA



(c) NASDAQ-100



(d) Gold and Euro

Figure 3. Radar maps showing pearson correlations of Twitter and SVI features vs commodities like oil (a); stock indices like DJIA (b) and NASDAQ-100 (c); forex rate of Euro and gold (d). (Blue and red correspond to significant correlation values in Figure 2)

for gold prices and Euro conversion rates. However important conclusions comes when we see behavior of Gold ETF's VIX, which is negative correlation prior one to two weeks; indicating increase gold related tweet discussions before dip in VIX index occurs. But it shows negative correlation at positive lag with search volumes. In contrast we observe a dip in VIX index (fear of buying gold) caused by increased discussion on Twitter as investors consider it as a safe investment, hence the confounding effect further observed in the search volumes.

Granger Causality Analysis

GCA rests on the assumption that if a variable X causes Y then changes in X will be systematically occur before the changes in Y. We realize lagged values of X shall bear significant correlation with Y. However correlation is not necessarily behind causation. Like the earlier approaches by [1, 9] we have made use of GCA to investigate whether one time series is significant in predicting another time series. GCA is used not to establish statistical causality, but as an economist tool to investigate a statistical pattern of lagged correlation. A similar observation that smoking causes lung cancer is widely accepted; proving it contains carcinogens but itself may not be actual causative of the real event i.e. cancer in this case.

Technique

Let returns R_t be reflective of fast movements in the stock market. To verify the change in returns with the change in Twitter features we compare the variance given by following linear models in equation 6 and equation 7.

$$R_t = \alpha + \Sigma^n{}_{i=1}\beta_i D_{t-i} + \epsilon_t \quad (6)$$

$$R_t = \alpha + \Sigma^n{}_{i=1}\beta_i D_{t-i} + \Sigma^n{}_{i=1}\gamma_i X_{i-t} + \varepsilon_t \quad (7)$$

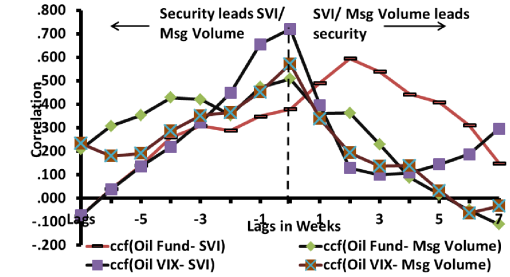
Equation 6 uses only 'n' lagged values of R_t , i.e. (R_{t-1}, . . .,R_{t-n}) for prediction, while equation 7 uses the n lagged values of both R_t and the tweet features time series given by $X_{t-1}, ..., X_{t-n}$. We have taken weekly time window to validate the casuality performance, hence the lag values [8] will be calculated over the weekly intervals $1, 2, ..., 7$.

Results

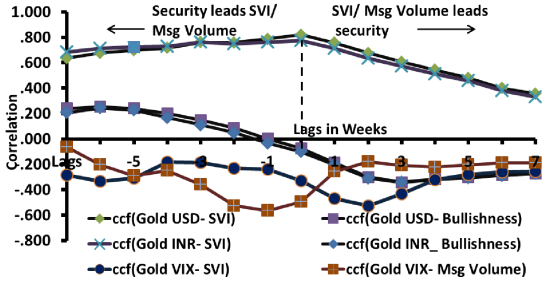
From the Table 3, we can reject the null hypothesis (H_o) that *the SVI and Twitter investor behavior do not affect returns in the financial markets* i.e. $\beta_{1,2,....,n} \neq 0$ with a high level of confidence (high p-values). However as we see the result applies to only specific negative and positive tweets (** for p-value < 0.05 and * for p-value < 0.1 which is 95% and 99% confidence interval respectively). Other features like agreement and message volume do not have significant casual relationship with the returns of a stock index (low p-values).

In Table 3 we can see that at the lag of one week, almost all the features are significant in predicting changes in the financial features of oil, DJIA, NASDAQ-100, gold and Euro. However as we go in the positive lag direction from 1st to 4 weeks, the significance decreases showing Twitter and SVI

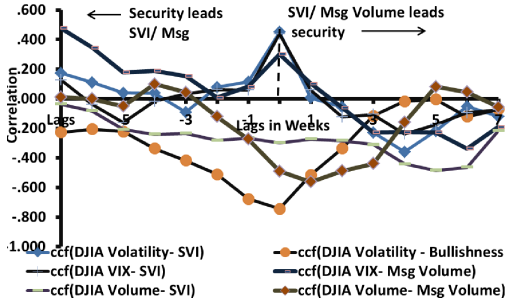
[8] *lag at k* for any parameter M at x_t week is the value of the parameter prior to x_{t-k} week. For example, value of returns for the month of April, at the lag of one month will be $return_{april-1}$ which will be $return_{march}$



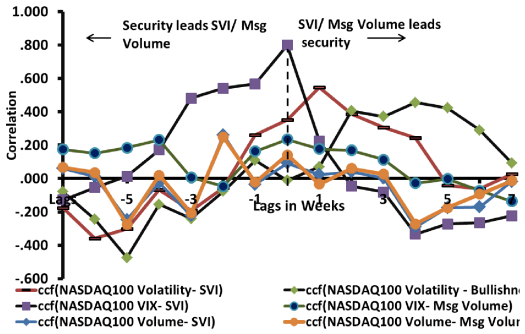
(a) Oil



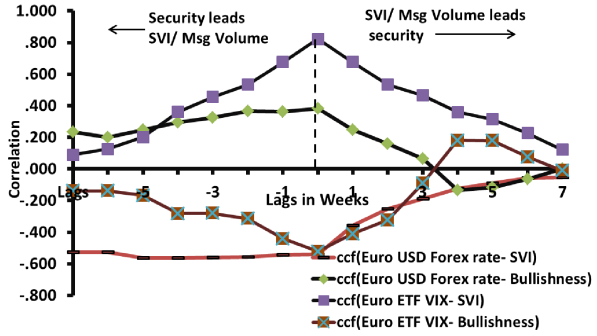
(b) GOLD



(c) DJIA



(d) NASDAQ-100



(e) EURO

Figure 4. Cross Correlation of Twitter and SVI features vs commodities like oil (a) and gold (b); stock indices like DJIA (c) and NASDAQ-100 (d); and forex rate of Euro (e)

mood series as Granger's causative of financial features. SVI shows uniform p values i.e. confidence of uptil 99% for almost all the sectors- both index (DJIA, NASDAQ-100) and commodities (gold, oil and forex rate of Euro). Twitter features specially for the indices- DJIA and NASDAQ-100 don't significance beyond 2-3 weeks, indicating the dispersive nature of information entropy on the social networks in contrast to the SVI factors.

EMMS model for Forecasting Analysis of Financial features

In this section we work upon the perennial question of *how much? and how good?* are these features proposed in the earlier sections can be useful to make accurate forecasts of financial indicators. For the same purpose we have used Expert Model Mining System (EMMS) which incorporates a set of competing methods such as Exponential Smoothing (ES), Auto Regressive Integrated Moving Average (ARIMA) and seasonal ARIMA models. These methods are widely used in financial modeling to predict the values of stocks/ bonds/ commodities/etc [15, 2]. These methods are suitable for constant level, additive trend or multiplicative trend and with either no seasonality, additive seasonality, or multiplicative seasonality.

Technique

Selection criterion for the EMMS is MAPE and stationary *R squared* which is measure of how good is the model under consideration as comapred to the baseline model [20]. The stationary R-squared can be negative with range $(-\infty, 1]$. A negative R-squared value means that the model under consideration is worse than the baseline model. Zero R-squared means that the model under consideration is as good or bad as the baseline model. Positive R-squared means that the model under consideration is better than the baseline model. Mean absolute percentage error (MAPE) is mean residuals (difference between fit value and observed value in percentage). To show the performance of tweet features in prediction model, we have applied the EMMS twice - first with SVI and Twitter sentiment features as independent predictor events and second time without them. This provides us with a quantitative comparison of improvement in the prediction using tweet features.

ARIMA (p,d,q) in theory and practice, are the most general class of models for forecasting a time series data, which is subsequently stationarized by series of transformation such as differencing or logging of the series Y_i. A stationary time series ΔY differences d times has stochastic component:

$$\Delta Y_i \hookrightarrow Normal(\mu_i, \sigma^2) \tag{8}$$

Where μ_i and σ^2 are the mean and variance of normal distribution, respectively. The systematic component is modeled as:

$$\mu_i = \alpha_i \Delta Y_{i-1} + ... + \alpha_p \Delta Y_{i-p} + \theta_i \varepsilon_{i-1} + ... + \theta_i \varepsilon_{i-q} \tag{9}$$

Where, ΔY the lag-p observations from the stationary time series with associated parameter vector α and ϵ_i the lagged errors of order q, with associated parameter vector.

Results

Table 3. Granger's Casuality Analysis- statistical significance (p values) at lags of 1,2,3 and 4 weeks between financial indicators and features of investor behavior (p - value $<$ 0.01:‡, p - value $<$ 0.05:†, p - value $<$ 0.1:*)

Securities		Lag	1	2	3	4
Oil	Close	Positive	**.009**†	**0.1***	0.755	0.238
		Negative	**.014**†	0.352	0.666	0.204
		Bull	**.014**†	0.25	0.77	0.238
		Msg Vol	**.018**†	**0.05**†	0.911	0.397
		Agreement	**0.061***	0.521	0.89	0.421
		SVI	**0.038**†	0.201	**0.006**†	**0.001**‡
	VIX	Positive	**0.048**†	0.966	0.454	0.746
		Negative	0.6	0.683	0.303	0.621
		Bull	**0.032***	0.819	0.364	0.742
		Msg Vol	**0.078***	0.701	0.706	0.949
		Agreement	**0.008**†	0.804	0.411	0.957
		SVI	**0.00002**‡	**0.001**‡	**0.037**†	**0.07***
DJIA	Return	Positive	0.675	0.601	0.986	0.266
		Negative	**0.065***	**0.056***	0.996	0.331
		Bull	0.38	0.442	0.991	0.305
		Msg Vol	**0.052***	0.608	0.947	0.237
		Agreement	0.264	0.243	0.826	0.552
		SVI	**0.021**†	**0.053***	**0.021**†	**0.057***
	VIX	Positive	0.461	0.501	0.936	0.683
		Negative	**0.024***	0.286	0.91	0.388
		Bull	0.38	0.527	0.672	0.583
		Msg Vol	**0.033***	**0.05***	0.666	0.97
		Agreement	0.427	0.436	0.616	0.752
		SVI	**0.015**†	**0.06***	**0.017**†	**0.03**†
Nasdaq	Return	Positive	**0.088***	**0.017**†	**0.049**†	**0.1***
		Negative	0.737	**0.017**†	**0.076***	**0.064***
		Bull	**0.061***	**0.024**†	0.213	0.136
		Msg Vol	0.253	0.218	0.043	0.473
		Agreement	**0.091***	0.31	0.988	0.245
		SVI	**0.091***	**0.081***	**0.064***	**0.091***
	VIX	Positive	**0.076***	**0.086***	**0.025**†	**0.042**†
		Negative	**0.001**‡	0.31	0.893	0.128
		Bull	**0.043**†	0.241	**0.021**†	**0.04**†
		Msg Vol	**0.179**	0.427	**0.024**†	0.148
		Agreement	**0.019**†	0.229	0.278	**0.093***
		SVI	**0.0002**‡	**0.002**‡	**0.02**†	**0.054***
Gold	Price USD	Positive	0.136	0.331	0.631	0.41
		Negative	0.56	0.712	0.807	0.66
		Bull	**0.004**‡	**0.023**†	**0.028**†	**0.058***
		Msg Vol	**0.027**†	**0.1***	0.625	0.557
		Agreement	**0.035**†	**0.009**‡	**0.015**†	**0.045**†
		SVI	**0.0001**‡	**0.00034**‡	**0.00041**‡	**0.001**‡
	VIX	Positive	**0.083***	**0.11***	0.192	**0.05**†
		Negative	**0.1***	0.454	**0.1***	**0.033**†
		Bull—	0.385	0.641	0.509	0.755
		Msg Vol	0.793	**0.1***	0.305	0.256
		Agreement	**0.1***	0.385	0.493	0.184
		SVI	0.414	**0.059***	**0.057***	**0.05**†
Euro	EURUSD	Positive	**0.051***	**0.11***	**0.1***	0.336
		Negative	**0.043***	0.51	0.249	0.561
		Bull	**0.069***	0.754	0.521	0.497
		Msg Vol	**0.1***	0.439	**0.1***	0.157
		Agreement	0.944	0.985	0.62	0.399
		SVI	**0.00001**‡	**0.00006**‡	**0.00008**‡	**0.0001**‡
	VIX	Positive	**0.1***	**0.085***	**0.092***	0.431
		Negative	**0.028**†	**0.011**†	**0.034**†	0.068
		Bull	0.498	**0.1***	**0.1***	0.797
		Msg Vol	0.443	0.256	0.987	0.213
		Agreement	0.384	0.587	0.55	0.557
		SVI	**0.091***	**0.0001**‡	**0.002**‡	**0.003**‡

Model equation for two cases are given below as equation 10 for forecasting without predictors and equation 11 for forecasting with predictors. In these equations Y is the financial feature- oil, gold, DJIA etc. and X represents the investor mood series from the SVI and Twitter features.

$$WithoutPredictors : Y_t = \alpha + \Sigma^n{}_{i=1}\beta_i Y_{t=i} + \epsilon_t \quad (10)$$

$$WithPredictors : Y_t = \alpha + \Sigma^n{}_{i=1}\beta_i Y_{t=i} + \Sigma^n{}_{i=1}\gamma_i X_{t=i} + \epsilon_t \quad (11)$$

Table 4. Forecasting results for the financial securities

Market Securities		Predictors	MAPE	Direction
US Oil Funds	Index	Yes	2.3202	75
		No	2.4203	62.5
	VIX	Yes	4.5592	75
		No	5.1218	56.3
DJIA	Index	Yes	0.8557	94.3
		No	1.1698	60
	VIX	Yes	5.3017	82.9
		No	5.6943	62.9
NASDAQ-100	Index	Yes	1.3235	90
		No	1.3585	50
	VIX	Yes	3.2415	83.3
		No	5.7268	50
Gold	USD	Yes	1.5245	78.6
		No	1.5555	64.3
	VIX	Yes	0.2534	71.9
		No	5.2724	56.1
Euro	EURUSD	Yes	2.6224	74.1
		No	4.3541	58.6
	VIX	Yes	4.4124	69
		No	4.7878	53.4

In the dataset we have time series for the total of 66 weeks, out which we use approximately 76% i.e. 50 weeks for the training both the models given in equation 11 and 10(for the time period 2nd June 2010 to 27th May 2011). Further we verify the model performance as one step ahead forecast over the testing period of 16 weeks from May 30th to 13 September 2011 which count for *wide and robust range of market conditions*. Forecasting accuracy in the testing period is compared for both the models in each case in terms of mean absolute percentage error (MAPE) and the direction accuracy. MAPE is given by the equation 12, where $\hat{y_i}$ is the predicted value and y_i is the actual value.

$$MAPE = \frac{\Sigma^n{}_i \left| \frac{y_i - \hat{y_i}}{y_i} \right|}{n} \times 100 \quad (12)$$

While direction accuracy is measure of how accurately market or commodity up/ down movement is predicted by the model, which is technically defined as logical values for $(y_{i,\hat{t}+1} - y_{i,t}) \times (y_{i,t+1} - y_{i,t}) > 0$ respectively. This is of prime importance to the high frequency traders and investors who hedge their investment in derivative markets as lots of prices (option premium, bonds etc.) are solely determined by the direction of the moving index or price.

As given in Table 4 we observe that the there is significant reduction in the values of MAPE for all the sectors for the forecasting model with the use of predictor sentiment and SVI series than the predictor model without the use of the these predictor series. Also for index values of DJIA direction accuracy of uptil 94.3% is achieved, while it is for 90% for NASDAQ-100. SVI and measure of wisdom of crowd

on Twitter gives quite a robust picture of how changing dynamics of the public opinion can be reflective of the market movements that would happen in near future.

DISCUSSIONS

As seen in Table 4, we observe one of the most significant improvements for NASDAQ-100's VIX (MAPE- 3.2415) and Gold VIX (MAPE- 0.2534), indicating the tech savvy investors who tweet a lot, hold significant power for the index movements. Comparing general prediction performance of behavior features (SVI + Twitter sentiment series) for market indices and commodity prices over the VIX index; better performance can be observed for VIX accounting for the fact that these behavior features are better indicative of investor fear before the actual price movement occurs in the stock. However for the forex price of Euro, investor sentiment is more centralized factor in controlling the price movement as compared to the VIX index. Results from the techniques-Grangers Causality Analysis and Cross-correlation (in Table 3 and figure 4) show that Twitter derived features affect the market instruments for a shorter duration of time, whereas search volume based features affect the market instruments for comparatively longer period of time. Our results show, there is no clear uniform pattern observed in the relationships across various elements of investor behavior and stocks over a wide spectrum of market securities and indices. Hence general conclusions regarding complexities about the performance cannot be drawn. We actively look forward into the future when fully automated bots will be advanced enough to understand all the behavioral mood dimensions of location specific discussions of what an investor is saying and make successful investments at minute time-scales.

CONCLUSION AND FUTURE WORK

Proposed approach combines the advantage of sophisticated statistical and linguistics summarization techniques. Such methods are able to capture a good picture of both the changing rates along with rise and fall probabilities for both commodities and stocks. Performance of the proposed model is more accurate in comparison to earlier works which were restricted only to mood states of entire Twitter feed applied in general to the market index [1, 9, 21]. We have made validation against larger tweet base, over bigger time period, with larger number of financial market instruments and greater prediction accuracy than any of the earlier works. Moreover, as far as practical implementation is concerned, our approach not only helps to improve index movements but also the present volatility and the VIX index which is the measure of the 30-days ahead market fear. More importantly it can be also used to determine portfolio adjustment decisions like ratio of risk to security for hedging with the greater confidence.

REFERENCES

1. Bollen, J., Mao, H., and Zeng, X.-J. Twitter mood predicts the stock market. *Computer 1010*, 3003v1 (2010), 1–8.
2. Box, G. E. P., and Jenkins, G. *Time Series Analysis, Forecasting and Control*. Holden-Day, Incorporated, 1990.
3. boyd, d. m., and Ellison, N. B. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication 13*, 1 (2007), 210–230.
4. Da, Z., Engelberg, J., and Gao, P. In search of attention. *Russell The Journal Of The Bertrand Russell Archives*, 919 (2010).
5. Das, S. R., and Chen, M. Y. Yahoo! for Amazon: Sentiment Parsing from Small Talk on the Web. *SSRN eLibrary* (2001).
6. Dewally, M. Internet investment advice: Investing with a rock of salt. *Financial Analysts Journal 59*, 4 (2003), 65–77.
7. Frank, M. Z., and Antweiler, W. Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *SSRN eLibrary* (2001).
8. Garman, M. B., and Klass, M. J. On the estimation of security price volatilities from historical data. *The Journal of Business 53*, 1 (1980), 67–78.
9. Gilbert, E., and Karahalios, K. Widespread worry and the stock market. *Artificial Intelligence* (2010), 58–65.
10. Go, A., Bhayani, R., and Huang, L.
11. Guresen, E., Kayakutlu, G., and Daim, T. U. Using artificial neural network models in stock market index prediction. *Expert Systems with Applications 38*, 8 (2011), 10389 – 10397.
12. Karabulut, Y. Can Facebook Predict Stock Market Activity? *SSRN eLibrary* (2011).
13. Mao, H., Counts, S., and Bollen, J. Predicting financial markets: Comparing survey,news, twitter and search engine data. Quantitative Finance Papers 1112.1051, arXiv.org, Dec. 2011.
14. Mao, H., Counts, S., and Bollen, J. Predicting Financial Markets: Comparing Survey,News, Twitter and Search Engine Data. *ArXiv e-prints* (Dec. 2011).
15. McCormick, G. P. Communications to the editorexponential forecasting: Some new variations. *Management Science 15*, 5 (1969), 311–320.
16. Miao, H., Ramchander, S., and Zumwalt, J. K. Information Driven Price Jumps and Trading Strategy: Evidence from Stock Index Futures. *SSRN eLibrary* (2011).
17. Nofsinger, J. R. Social mood and financial economics. *Journal of Behavioral Finance 6*, 3 (2005), 144–160.
18. Qian, B., and Rasheed, K. Stock market prediction with multiple classifiers. *Applied Intelligence 26* (February 2007), 25–33.
19. Ruiz, E. J., Hristidis, V., Castillo, C., Gionis, A., and Jaimes, A. Correlating financial time series with micro-blogging activity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, ACM (New York, NY, USA, 2012), 513–522.

Table 7. Vector Matrix for Gold SVI Factors (*Only 1 factor for gold as search terms for gold fall on the same dimension plane in the feature vector map.*)

Search Term	Gold SVI factors Fact 1
buy gold	.575
invest in gold US data	.942
invest in gold worldwide	.884
dollar to pound exchange rate	.905
dollar to pound exchange	.904

Table 8. Vector Matrix for Nasdaq SVI Factors

Search Terms	Nasdaq SVI Factors Fact 1	Fact 2
nasdaq today	.833	.398
dow futures quote	.888	.239
futures quote	.673	-.004
NASDAQ quote	.821	.326
nylc	.257	.936
bank of america dividends	.161	.947

Table 5. Vector Matrix for Oil SVI Factors

Search Terms	Oil SVI Factors Fact 1	Fact 2
oil commodity	.671	.688
crude oil etf	.319	.897
oil funds	.304	.900
oil etf	.303	.898
oil quotes	.593	.599
oil prices per barrel	.853	.338
spot oil prices	.727	.533
wti crude	.361	.504
how much oil is left	.452	.446
futures price	.521	.759
how to buy oil	.298	.918
oil ticker	.634	.655
current oil	.741	.021
crude oil futures	.441	.639
crude oil price	.442	.700

Table 6. Vector Matrix for DJIA SVI Factors

Search Terms	Dow SVI Factors Fact 1	Fact 2
djia	.931	.038
dow jones industrial average	.811	.495
dow jones	.929	.116
dow	.966	.010
s&p 500	.635	-.032
Stock Market	.689	-.337
stock message board	-.012	.936

20. Sabani, L. Forecasting structural time series models and the kalman filter, a. c. harvey. cambridge university press, cambridge, 1989. isbn 0-521-32196-4, cloth, 55.00 pp. xvi + 554. *Journal of Applied Econometrics 6*, 3 (1991), 329–331.

21. Sprenger, T. O., and Welpe, I. M. Tweets and Trades: The Information Content of Stock Microblogs. *SSRN eLibrary* (2010).

22. Tversky, A., and Kahneman, D. Prospect theory: An analysis of decision under risk. Tech. rep., 1979.

23. Zhang, X., Fuehres, H., and Gloor, P. A. Predicting stock market indicators through twitter i hope it is not as bad as i fear. *Anxiety* (2009), 1–8.

APPENDIX

The PCA component matrixes for Oil, DJIA, NASDAQ-100, Gold and Euro are given in Tables 5, 6, 8, 7 and 9 respectively. Effect of so many search terms can be concisely mapped to double or single factors i.e. original high-dimensional data onto a lower dimensional space. The new PCA factors are uncorrelated, and are ordered by the fraction of the total information each retains and filtered out on the basis of Kaiser criterion, with threshold eigen value greater than 1.

Table 9. Vector Matrix for Euro SVI Factors

Search Terms	Euro SVI factors Fact 1	Fact 2
exchange rates converter	.530	-.383
dollar euro exchange rate history	.798	.079
rupee exchange rate	-.047	.063
oanda currency	.105	.810
rupee exchange	.929	-.102
dollar rupee exchange rate	.938	.043
bloomberg live tv	.828	-.228
eurusd	-.237	.741